AFRL-HE-WP-TP-2006-0026

AIR FORCE RESEARCH LABORATORY

Speaker Segmentation and Clustering Using Gender Information

Brian M. Ore

General Dynamics Advanced Information Solutions
Dayton OH

Raymond E. Slyh
Eric G. Hansen

Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433-7022

February 2006

Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
February 2006	Proceedings	

4. TITLE AND SUBTITLE
Speaker Segmentation and Clustering Using Gender Information

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
*Brian M. Ore, **Raymond E. Slyh, **Eric G. Hansen

5d. PROJECT NUMBER
7184

5e. TASK NUMBER
08

5f. WORK UNIT NUMBER
71

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)
*General Dynamics Advanced Information Systems
Dayton OH

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
**Air Force Materiel Command
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433-7022

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/HECP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-WP-TP-2006-0026

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
This will be published in the Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop. The clearance numbers are AFMC/PAX-06-068, AFRL/WS-06-0429, cleared 17 February 2006.

14. ABSTRACT
This paper considers the segmentation and clustering of conversational speech for the two-wire training (3conv2w) and two-wire testing (1conv2w) conditions of the NIST 2005 Speaker Recognition Evaluation. A notable feature of the system described is that each file is labeled as containing either opposite- or same-gender speakers. The speech segments for opposite-gender files are clustered by gender, while those for same-gender files are processed by agglomerative clustering. By using gender information in the clustering of the opposite-gender files, the equal error rate in the 3conv2w training condition was reduced from 15.2% to 9.9%. For the 1conv2w testing condition, clustering opposite-gender files by gender did not improve performance over agglomerative clustering; however, it was over 100 times faster than agglomerative clustering on the opposite-gender files.

15. SUBJECT TERMS
Speaker segmentation, clustering, gender

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Raymond Slyh
UNC	UNC	UNC	SAR	10	19b. TELEPHONE NUMBER *(include area code)* (937) 255-9248

Speaker Segmentation and Clustering using Gender Information

Brian M. Ore,[1] *Raymond E. Slyh,*[2] *and Eric G. Hansen*[2]

[1]General Dynamics Advanced Information Systems, Dayton OH, USA
[2]Air Force Research Laboratory, Human Effectiveness Directorate,Wright-Patterson AFB OH, USA

Abstract

This paper considers the segmentation and clustering of conversational speech for the two-wire training (3conv2w) and two-wire testing (1conv2w) conditions of the NIST 2005 Speaker Recognition Evaluation. A notable feature of the system described is that each file is labeled as containing either opposite- or same-gender speakers. The speech segments for opposite-gender files are clustered by gender, while those for same-gender files are processed by agglomerative clustering. By using gender information in the clustering of the opposite-gender files, the equal error rate in the 3conv2w training condition was reduced from 15.2% to 9.9%. For the 1conv2w testing condition, clustering opposite-gender files by gender did not improve performance over agglomerative clustering; however, it was over 100 times faster than agglomerative clustering on the opposite-gender files.

1. Introduction

Segmentation and clustering of speech into homogeneous segments is desirable for a variety of speech and speaker recognition applications. For example, in applications such as broadcast news or meeting transcription, improved speech recognition rates can be obtained by first segmenting an audio recording into homogeneous segments based on gender and channel and then using gender- and channel-dependent models in the decoding procedure [1, 2]. In addition, it is useful to be able to retrieve indexed broadcast news or meeting segments based on speaker identity [2]. In this paper, we consider the problem of speaker segmentation and clustering in the context of the annual NIST Speaker Recognition Evaluation (SRE) [3]. Specifically, we consider the two-wire training and testing conditions, in which both sides of the conversation are summed to give a single-channel file.

For the purposes of this paper, segmentation is defined as finding the time marks in an audio recording where there is a change in speaker identity. One of the simplest methods for accomplishing this is to declare a change wherever there is a speech/non-speech boundary [2]. A more sophisticated method, proposed by Chen and Gopalakrishnan [4], is to use the Bayesian Information Criterion (BIC) to detect change points. This method is based on modeling the audio stream as a Gaussian process and using a maximum likelihood approach to detect potential changes. There have been numerous adaptations of the BIC procedure, including the DISTBIC procedure of [2] and the modified BIC procedure of [5] that removes the necessity for having to choose the threshold in the penalty term. In [6, 7], gender and channel detection have been used in the first stages of segmentation for speaker diarization of news broadcasts.

Clustering can be defined as grouping homogeneous speech segments. A common approach to accomplish this is to model each speech segment as either a single Gaussian [4] or as a Gaussian Mixture Model (GMM) [8, 9] and to use hierarchical agglomerative clustering to combine the segments. If the number of data classes and speakers is known, then the clustering procedure can be terminated when the number of clusters reaches a predefined limit.

In this paper, we describe the speaker segmentation and clustering system that we submitted for the two-wire training and testing conditions of the NIST 2005 SRE. A notable feature of the system is that each file is labeled as containing either opposite- or same-gender speakers. The speech segments for opposite-gender files are clustered by gender, while those for same-gender files are processed by agglomerative clustering. We show that using gender information in the clustering of the opposite-gender files improves the performance in the two-wire training condition, while no improvement was seen on the two-wire testing condition. However, clustering by gender was still useful in the two-wire testing condition as it was over 100 times faster than using agglomerative clustering on the opposite-gender files.

In addition to describing the submitted system and showing its performance, we show the results of two experiments conducted after the official evaluation. The first experiment examined the utility of adding a speaker change detector (SCD) based on the modified BIC of [5], while the second experiment examined the utility of using a different metric to determine which segments to cluster–namely, the likelihood ratio (LR) used in [8, 9].

This paper is organized as follows. The next section discusses the NIST SRE two-wire training and testing

Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Air Force.

conditions. Section 3 discusses the system components used to perform the segmentation and clustering. Section 4 shows the performance of the system on the NIST 2005 two-wire training and testing conditions, while Section 5 shows the results of the two post-evaluation experiments. Section 6 presents the conclusions.

2. The NIST 2005 SRE

The NIST 2005 SRE consisted of 20 distinct tasks, which involved various combinations of training and testing conditions [3]. The various conditions were generally delineated based on the amount of data provided and on whether the data files were four-wire or two-wire files. The four-wire files consisted of both sides of telephone conversations, where each side was on a separate channel and the channel of interest was designated. The two-wire files consisted of a single channel containing the sample-by-sample sum of the two conversation sides. Thus, two-wire files needed to be segmented and clustered. For the purposes of this paper, only a subset of the various training and testing conditions is considered, and for these conditions, each file was of approximately five minutes in duration.

The training conditions considered are designated as "1conv4w," "3conv4w," and "3conv2w." For the 1conv4w training condition, each speaker model was to be built using a single four-wire (two-channel) file, with the target speaker channel designated. For the 3conv4w training condition, each speaker model was to be built from three four-wire files, each involving the target speaker on their designated sides. For the 3conv2w training condition, each speaker model was to be built from three two-wire (single channel) files, where each file contained a conversation between the target speaker and another speaker. The target speaker was the same across all three files, while the non-target speakers were all distinct. Note that in all of the training conditions, the gender of the target speaker was known.

The testing conditions considered are designated as "1conv4w" and "1conv2w." The 1conv4w testing condition consisted of individual test files similar to those in the 1conv4w training condition. The 1conv2w testing condition consisted of individual two-wire files, and the task was to determine if the target speaker was one of the speakers in the file.

We designate a particular task by its training designator followed by its testing designator. Thus, 3conv2w-1conv4w would denote the 3conv2w training condition with the 1conv4w testing condition.

It is important to note that, while the 3conv2w training condition is the two-wire analog of the 3conv4w training condition, there is not an exact one-for-one correspondence between the models in the two conditions. Thus, the 3conv4w-1conv4w task gives an *idea* of what the performance might be for the 3conv2w-1conv4w task if one had a perfect segmentation and clustering system, but it is not an exact bound. Likewise, for the 1conv2w and 1conv4w testing conditions.

The data files contained calls from both cellular and land-line handsets. Most conversations were in English, but there were some conversations in Arabic, Mandarin, Russian, and Spanish.

3. System Description

This section describes the various components of the system. An overview of the system is as follows. The first step of the system is to segment each file into speech and non-speech regions with a speech activity detector (SAD). The second step uses a gender detector to label the gender of each speech segment in each file. The third step is to determine if each file contains same-gender speakers or opposite-gender speakers. If a file contains opposite-gender speakers, it is clustered by the gender labels of the speech segments. If a file contains same-gender speakers, it is clustered by an agglomerative clustering method. For the 3conv2w training condition, the clusters are used in building speaker models. For the 1conv2w testing condition, the clusters are each scored against the target models, and the maximum score is taken as the score of the test file.

The next subsection describes the data used for developing several components of the system, and Subsection 3.2 describes the SAD. Subsection 3.3 describes the gender-based segmentation and clustering procedure, while Subsection 3.4 describes the agglomerative clustering procedure. Finally, Subsection 3.5 describes the process for building speaker models.

3.1. Development Data

A number of components of the system were built using the same development data—namely, the SAD, the gender detector, the background model, and the channel/gender models used in feature mapping [10]. This development data consisted of approximately 16 hours of speech and 13 hours of non-speech. Data were selected from the OGI National Cellular database[1] (for analog cellular data) and from the NIST 2001–2003 SREs (for land-line electret, land-line carbon button, and digital cellular data). The data were balanced by gender and channel, and the audio files consisted of only English speech.

Some system parameters were tuned using the data from the two-wire tasks of the NIST 2004 SRE.

3.2. Speech Activity Detection

The SAD worked in three stages. The first stage utilized a two-state speech/non-speech Hidden Markov Model (HMM) to define the initial speech/non-speech bound-

[1] See: http://cslu.cse.ogi.edu/corpora/corpCurrent.html

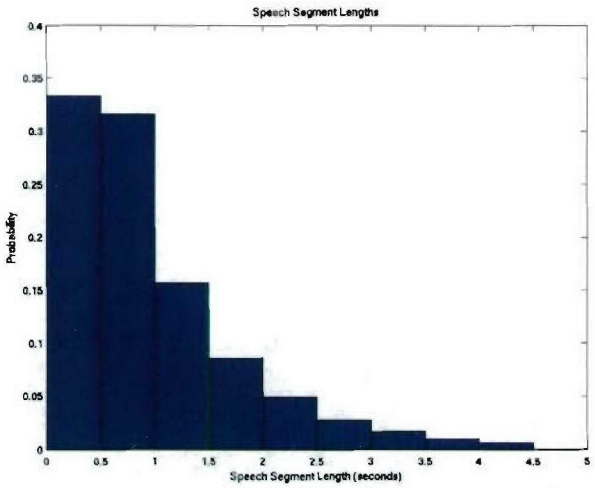

Figure 1: Histogram of speech segment lengths after speech activity detection computed over the 3conv2w training files of the NIST 2005 evaluation.

aries. The HMM was trained with HTK[2] using 64 mixtures per state and diagonal covariance matrices. The feature set consisted of 19 mel-frequency cepstral coefficients (MFCCs) bandlimited to 300–3138 Hz with the 0^{th} coefficient removed. Deltas of the features were included, but no channel compensation was applied. The HMM was built using the development data mentioned in the previous section, where the "truth" labels for the speech and non-speech segments used in training the HMM came from the start and end times of the words output by the SONIC speech recognition system [11, 12].

The second stage refined the HMM output by applying an energy-based detector—namely, the *xtalk* program from version 2.1 of the MIT Lincoln Laboratory (MIT-LL) MFCC/GMM speaker recognition system [13]. An energy threshold was estimated from the ten non-speech segments output from the HMM classifier with the highest likelihood. The average energy of these segments was used as the threshold.

The final stage post-processed the output by reclassifying as non-speech any segments labeled as speech that were less than 20 msec in duration. This stage helped to remove spurious noise segments.

Figure 1 shows a histogram of the resulting speech segment lengths after the SAD was applied to the files used in the NIST 2005 3conv2w training condition. One can see that the speech segments tended to be rather short with 64.3% of the segments having a length of 0–1 sec, 23.9% of the segments having a length of 1–2 sec, and 7.6% of the segments having a length of 2–3 sec. These short segments helped to obviate the need for speaker change detection, as is shown in Section 5.1.

[2] Available at: http://htk.eng.cam.ac.uk/

3.3. Gender-Based Segmentation and Clustering

The first step of this procedure was to label each speech segment from the SAD as either male or female using an MFCC/GMM-based gender detector. A gender-independent GMM was built using the MIT-LL MFCC/GMM system with the data described in Subsection 3.1; this GMM used 2048 mixtures and diagonal covariance matrices. Next, male and female GMMs were built by adapting the mixture means of the gender-independent GMM using MAP adaptation. The feature set used was the same as in Subsection 3.2, except that RASTA filtering [14] was applied to compensate for channel effects. Each speech segment was scored using the male and female GMMs. For the files used in the 1conv2w testing condition, the gender label for each segment was assigned based on which gender model scored higher for the segment.

For the files used in the 3conv2w training condition, the gender of the target speaker was known, and this fact was used to bias the gender decision for each segment. If the target speaker was male, then for a speech segment to be classified as female, it had to score better against the female model than it scored against the male model by at least a threshold. A similar procedure was used when the target speaker was female. The threshold values were determined using the NIST 2004 SRE data and found to be language-specific, ranging from 0.01 for Arabic files to 0.12 for Mandarin files.

Once the speech segments for a file were labeled by gender, the file was determined to be either from same- or opposite-gender speakers. This was done based on the number of frames labeled as each gender. If less than N percent of the total speech frames were classified as the same gender, then the file was classified as containing opposite-gender speakers; otherwise, it was classified as containing same-gender speakers. The value of N was also language-dependent and determined from the NIST 2004 SRE data, ranging from 0.90 for Arabic to 0.94 for Mandarin and Russian.

Table 1 shows the actual and estimated percentages of the number of opposite-gender files per model in the NIST 2004 3conv training condition (the 2004 analog of the 2005 3conv2w training condition), while Table 2 shows the actual and estimated percentages of the number of opposite-gender files per model in the NIST 2005 3conv2w training condition. One can see that the procedure for determining if a file was of same or opposite genders performed reasonably well. It is interesting to note that the 2005 3conv2w training condition only had 20.4% of the models being built using files containing only same-gender speakers; thus, there was considerable potential for using gender-based segmentation and clustering in the 2005 SRE.

It is important to note that even if a file was misclassified as to its status as having same- or opposite-

Number of Opposite-Gender Files/Model	Actual Percentage	Estimated Percentage
0	39.4	31.4
1	22.1	21.6
2	17.7	24.5
3	20.8	22.5

Table 1: Actual and estimated percentages for the number of opposite-gender files per model in the NIST 2004 3conv training condition.

Number of Opposite-Gender Files/Model	Actual Percentage	Estimated Percentage
0	20.4	19.1
1	37.0	32.2
2	30.5	33.6
3	12.1	15.1

Table 2: Actual and estimated percentages for the number of opposite-gender files per model in the NIST 2005 3conv2w training condition.

gender speakers, it does not mean that the file was not well segmented and clustered. If both speakers were of the same gender, it is entirely possible that one speaker scored higher against the female GMM while the other scored higher against the male GMM. On the other hand, if the speakers were of opposite gender and the file were mis-classified as having same-gender speakers, the speech segments may still have been properly segmented and clustered using the agglomerative clustering procedure of the next subsection.

3.4. Agglomerative Clustering

The procedure used to cluster similar speech segments is illustrated in Figure 2. The first step in the clustering algorithm was the initialization of the models used to represent each segment. This was accomplished by first training a GMM using all of the speech segments in a file, and then using MAP adaptation to adapt the weights of this model to fit the characteristics of each segment [8, 13], thus creating a separate model for each speech segment. The GMMs each consisted of 64 mixtures and used diagonal covariance matrices. The feature set was the same as that discussed in Section 3.2, except that the MFCCs were bandlimited to 200–2860 Hz. This bandwidth for the MFCCs was found to perform better in our agglomerative clustering experiments done use data from past NIST SREs.

Next, the feature vectors for each of the speech segments were scored against all of the models for the other segments. The highest scoring pair of segments were clustered together. The feature vectors of the clustered segments were concatenated and a new model was estimated by averaging the weights of the two models. This process was repeated until there were three remaining segments. The final number of segments was chosen as three to account for speech segments that contain multiple speakers due to a missed speaker change point or segments not containing actual speech due to errors by the SAD. Presumably, there would be one cluster for each of the two speakers and a third "garbage" cluster. In prior work on the two-wire conditions of the NIST 2002 SRE, we found no performance difference between stopping at two clusters and stopping at three clusters.

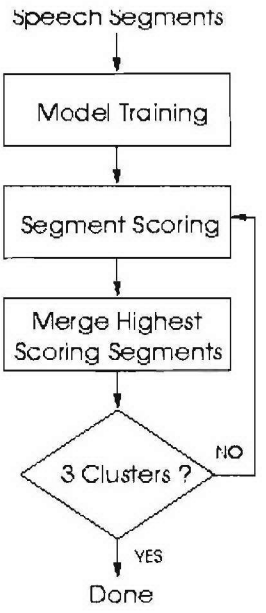

Figure 2: Block diagram of the agglomerative clustering algorithm.

3.5. Building Speaker Models

For the 3conv2w training condition, an additional step needed to be performed once each file had been segmented and clustered—namely, to determine which clusters from the three training files for a target speaker to include in the final set of clusters used to build the model. This procedure varied, depending on how the various files were clustered.

If none of the three files for a target speaker were segmented and clustered by gender, then the procedure was as follows. After agglomerative clustering, there were a total of nine clusters with three from each conversation. The agglomerative clustering algorithm was used to find three final clusters across the three training files representing the common speaker, with the restriction that only one cluster from each file could be used.

If at least one of the files was segmented and clustered by gender, then the procedure for clustering across the three files was as follows. First, for each file seg-

mented and clustered by gender, the top 90% of the segments were taken in which the given target speaker gender scored highest. (Only the top 90% of the segments were used to help compensate for possible segmentation and gender classification errors.) The simplest training task was when all three speech files were of opposite gender. The target speaker clusters were extracted from each file independently based on gender and all of the extracted speech frames were pooled to train a speaker model. In the situation where one or two of the training files was/were classified as opposite-gender, the target speaker clusters were extracted from the opposite-gender files and used to train an initial seed model. The agglomerative clustering algorithm was then used across all of the same-gender training files, with the restriction that the seed model was always one of the merged models.

Version 2.1 of the MIT-LL MFCC/GMM system [13] was used to perform the model training and the testing for all experiments. The feature set consisted of 19 MFCCs computed every 10 msec and bandlimited to 300–3138 Hz with the 0^{th} coefficient removed and deltas of the features included. RASTA filtering [14], feature mapping [10], and mean and variance normalization were applied for channel compensation. The background model and the channel/gender models used in feature mapping used 2048 mixtures and diagonal covariance matrices and were built using the data described in section 3.1. Gender-dependent T-norm [15] was applied in all experiments using 120 models for each gender, where each model was trained with approximately two minutes of speech gathered from the NIST 2001–2003 SREs. In building target and T-norm models, only the mixture means were adapted from those of the background model using MAP adaptation [13].

For the 1conv4w and 3conv4w training conditions, all frames labeled as speech by the SAD were used to build target models with the MFCC/GMM procedure just described.

4. System Performance

This section shows the performance of the segmentation and clustering system for some of the NIST 2005 two-wire tasks. Figure 3 shows the performance of the MFCC/GMM system on (1) the 3conv2w-1conv4w task with and without gender-based segmentation and clustering and (2) the 3conv4w-1conv4w task. On the 3conv2w-1conv4w task, the system yields an equal error rate (EER) of 9.9% with gender-based segmentation and clustering and an EER of 15.2% without gender-based segmentation and clustering. The performance of both clustering systems on the 3conv2w task lags that of the baseline system on the 3conv4w task, which yields an EER of 7.1%.

Figure 4 shows the performance of the MFCC/GMM system on (1) the 1conv4w-1conv2w task with and without gender-based segmentation and clustering and (2) the 1conv4w-1conv4w task. On the 1conv4w-1conv2w task, the system yields an EER of 11.7% with gender-based segmentation and clustering and an EER of 11.2% without gender-based segmentation and clustering. The performance of both clustering systems on the 1conv2w task again lags that of the baseline system on the 1conv4w task, which yields an EER of 10.3%.

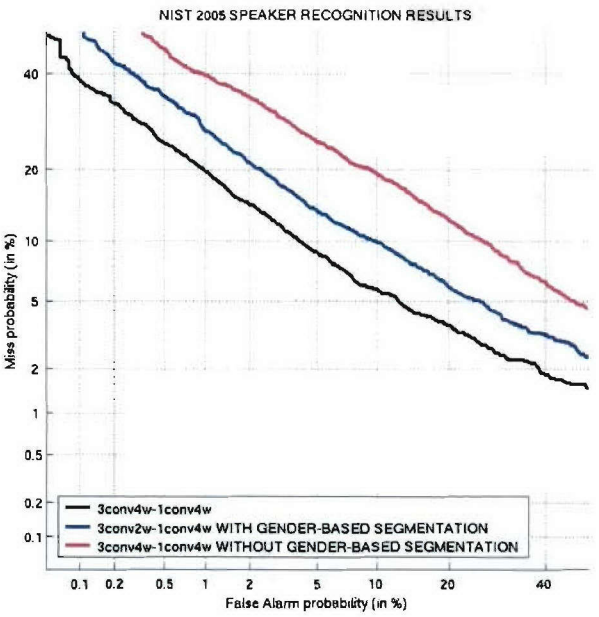

Figure 3: Performance of the MFCC/GMM system on (1) the 3conv2w-1conv4w task with and without gender-based segmentation and clustering and (2) the 3conv4w-1conv4w task.

Figure 5 shows the system performance over all possible combinations of the 3conv4w and 3conv2w training conditions with the 1conv4w and 1conv2w testing conditions, where the system used gender-based segmentation and clustering for the two-wire conditions. It is interesting to note that there is little difference in performance for the 3conv2w-1conv4w and 3conv4w-1conv2w tasks.

5. Post-Evaluation Experiments

This section describes results of two experiments that were conducted after the official 2005 evaluation. The first experiment evaluated the utility of adding speaker change detection based on the BIC technique of [5]; this experiment is discussed in Subsection 5.1. The second experiment evaluated the utility of changing the metric used in the clustering procedure from one based on scoring each segment against every other segment's model and clustering based on the highest score to one based on the LR described in [8]; this experiment is discussed in Subsection 5.2. Finally, Subsection 5.3 looks at the computational burden of the clustering techniques.

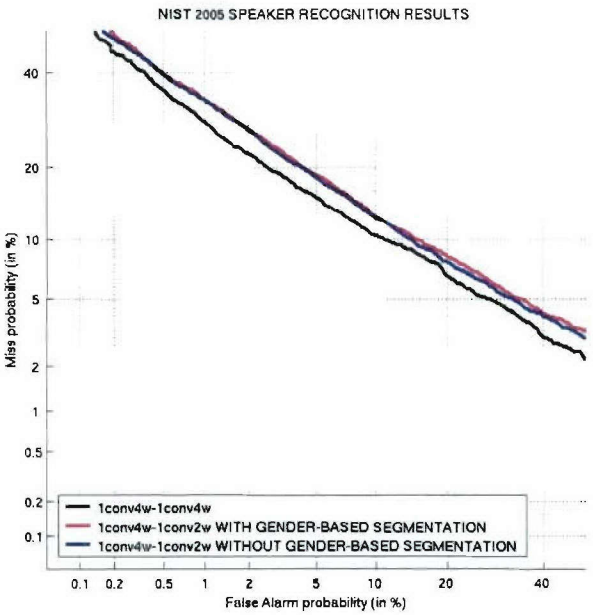

Figure 4: Performance of the MFCC/GMM system on (1) the 1conv4w-1conv2w task with and without gender-based segmentation and clustering and (2) the 1conv4w-1conv4w task.

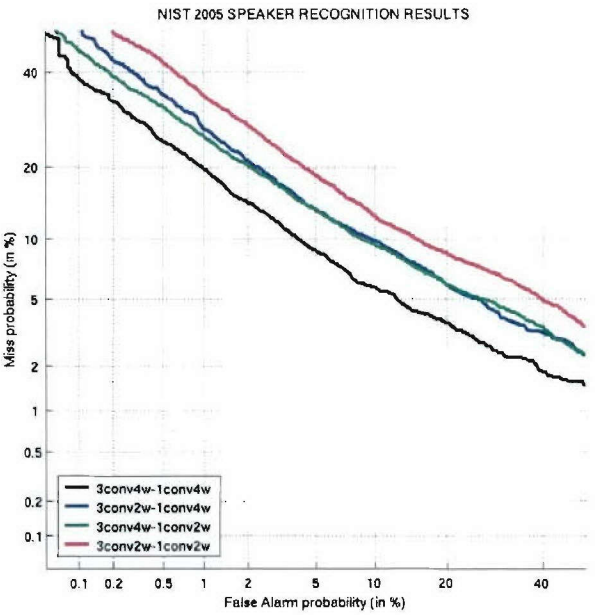

Figure 5: MFCC/GMM system performance over the 3conv4w and 3conv2w training and 1conv4w and 1conv2w testing conditions.

5.1. Effect of BIC-Based Speaker Change Detection

The modified BIC procedure of [5] is as follows. Consider a segment, Z, with a point, t, splitting the segment into two adjacent segments. Let X denote the portion of Z up to time t, and let N_X denote the number of feature vectors in X. Let Y denote the portion of Z after time t, and let N_Y denote the number of feature vectors in Y. Under the null hypothesis, H_0, that t does not constitute a speaker change point, model Z as a GMM with two mixtures, and calculate the parameters, θ_Z, of this GMM with the expectation-maximization algorithm. The log likelihood of Z under H_0 is

$$L_0 = \sum_{i=1}^{N_X} \log p(x_i|\theta_Z) + \sum_{i=1}^{N_Y} \log p(y_i|\theta_Z).$$

Let H_1 denote the hypothesis that t is a speaker change point. In this case, model the data in X with a single Gaussian density having parameters θ_X, and model the data in Y as a single Gaussian density having parameters θ_Y. Then the log likelihood under H_1 is

$$L_1 = \sum_{i=1}^{N_X} \log p(x_i|\theta_X) + \sum_{i=1}^{N_Y} \log p(y_i|\theta_Y).$$

The difference between this model and that of the standard BIC is that, under H_0, Z is modeled here using the GMM with two mixtures, and X and Y are modeled as single Gaussian densities, whereas X, Y, and Z are all modeled as single Gaussian densities in the standard BIC. Now, let $d_t = L_1 - L_0$. If d_t is a local maxima greater than zero, declare t to be a speaker change point.

This BIC-based speaker change detector (SCD) was investigated for further refining the segments output by the SAD. Figure 6 shows the effect on the system detection performance of using this SCD. One can see that the BIC-based SCD degrades the performance regardless of whether the gender-based segmentation is used. Given that 88.2% of the speech segments output by the speech activity detector are no longer than two seconds in duration, it would appear that there is no benefit to be gained by using the BIC-based SCD with the current segmentation and clustering system.

5.2. Effect of Likelihood Ratio Clustering

To determine which segments to cluster in the agglomerative clustering procedure, the submitted system scored segments against models built for all of the other segments and clustered the segments with the highest scoring pair of data and model, we refer to this procedure as score-based clustering. This experiment examined the use of the LR of [8] for determining which segments to cluster. Consider two segments of feature vectors, X and Y, and the union of the feature vectors from these two segments, Z. Let $L(X|\theta_X)$ be the likelihood of X, where θ_X denotes the maximum likelihood estimate for the parameters of a GMM trained with the vectors in X. Define $L(Y|\theta_Y)$ and $L(Z|\theta_Z)$ similarly. The likelihood that segments X and Y were generated by differ-

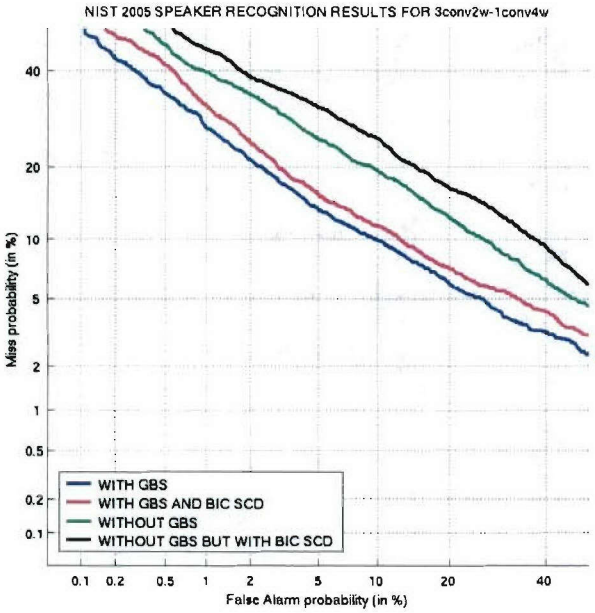

Figure 6: Performance of the segmentation and clustering system on the 3conv2w-1conv4w task (1) with gender-based segmentation (GBS), but without the SCD (denoted "WITH GBS"); (2) with GBS and the SCD (denoted "WITH GBS AND BIC SCD"); (3) without both GBS and the SCD (denoted "WITHOUT GBS"); and (4) without GBS, but with the SCD (denoted "WITHOUT GBS BUT WITH BIC SCD").

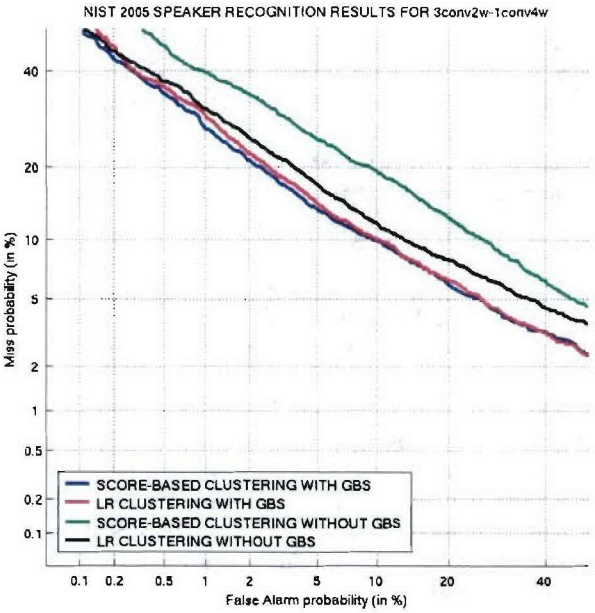

Figure 7: Performance of the segmentation and clustering system on the 3conv2w-1conv4w task using (1) score-based clustering with gender-based segmentation (GBS), (2) LR clustering with GBS, (3) score-based clustering without GBS, and (4) LR clustering without GBS.

ent speakers is $L_1 = L(X|\theta_X)L(Y|\theta_Y)$. The likelihood that the segments were generated by the same speaker is $L_0 = L(Z|\theta_Z)$. Define the LR as $\lambda = L_0/L_1$. For each pass of the agglomerative clustering system, combine the two segments with the highest LR.

Figure 7 shows the effect on the system performance of using the LR as the metric for determining which segments to cluster at each stage. The LR clustering method performs better than the original GMM scoring-based metric when no gender-based segmentation is used. Using gender-based segmentation improves the system performance for both metrics and yields slightly better performance when coupled with the GMM scoring-based metric than it does with the likelihood ratio metric, although the difference between the two is slight.

5.3. Computational Burden

To partially investigate the computational burden of the gender-based segmentation and clustering versus that of the agglomerative clustering using both the score-based metric as well as the LR, 100 opposite-gender files from the 2005 SRE were processed using the three clustering methods. The gender-based segmentation and clustering took 42 minutes to process the files, the score-based clustering method took 5688 minutes, and the LR clustering method took 4815 minutes. Thus, the gender-based clustering was 135 times faster than the score-based clustering and 114 times faster than the LR clustering.

6. Discussion and Conclusions

In this paper, we presented a segmentation and clustering algorithm for the two-wire conditions of the NIST 2005 SRE. Incorporating gender information in the segmentation and clustering yielded a significant improvement in the performance of our system for the 3conv2w-1conv4w task, reducing the EER from 15.2% to 9.9%. The baseline MFCC/GMM system yielded an EER of 7.1% on the 3conv4w-1conv4w task (which does not require segmentation and clustering), indicating that additional improvement in segmentation and clustering performance is still desirable. For the 1conv4w-1conv2w task, including gender information was found to have an insignificant effect on the system performance, but it still provided a significant reduction in the computational burden by over a factor of 100.

Additional research is still needed to determine why the gender-based segmentation and clustering did not improve the performance for the 1conv4w-1conv2w task, whereas it did improve the performance for the 3conv2w-1conv4w task. One possibility is that the performance difference may be linked to the biasing of the gender detector scores toward the known target speaker gender for

the 3conv2w-1conv4w task, whereas the gender detector scores were not biased in the 1conv4w-1conv2w task due to the rules of the evaluation.[3] The underlying gender detector performance may need to be improved without the need for biasing the scores in order for additional improvement to be obtained in the 1conv4w-1conv2w task.

The post-evaluation experiments yielded some interesting results. The incorporation of the SCD based on the modified BIC of [5] degraded the performance of the system for the 3conv2w-1conv4w task, regardless of whether gender-based segmentation and clustering was used or not. The initial speech segments generated by the SAD are already short, and the SCD only served to break some of them up into yet smaller segments. These smaller segments can lead to additional errors in gender detection and in agglomerative clustering. It would appear that these additional errors outweigh any potential errors caused by missing a speaker change point.

The second post-evaluation experiment showed that the use of the LR metric of [8] in the agglomerative clustering improved the performance of the system on the 3conv2w-1conv4w task compared to the original score-based metric as long as gender-based segmentation and clustering was not used. However, when the gender-based segmentation and clustering was used, this improvement vanished, and the score-based metric outperformed the LR metric slightly in the typical NIST operating region. Regardless of the metric used in the agglomerative clustering, the gender-based segmentation and clustering improved the performance on this task. Future research will investigate the use of other metrics in the clustering.

7. References

[1] P. Woodland, "The development of the HTK broadcast news transcription system: An overview," *Speech Comm.*, vol. 37, pp. 47–67, May 2002.

[2] P. Delacourt and C. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communication*, vol. 32, pp. 111–126, September 2000.

[3] NIST, *The NIST Year 2005 Speaker Recognition Evaluation Plan*, Version 6, 29 March 2005. (Available at: http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.pdf)

[4] S. Chen and P. Gopalakrishnan, "Speaker, environment and change detection and clustering via the Bayesian information criterion," *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, (Lansdowne VA), Feb. 1998.

[5] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 649–651, August 2004.

[6] S. Meignier, D. Moraru, C. Fredouille, L. Besacier, and J.-F. Bonastre, "Benefits of prior acoustic segmentation for automatic speaker segmentation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Montreal, Quebec, Canada), May 2004.

[7] D. Istrate, N. Scheffer, C. Fredouille, and J.-F. Bonastre, "Broadcast news speaker tracking for ESTER 2005 campaign," in *Proc. of INTERSPEECH*, (Lisbon, Portugal), September 2005.

[8] L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian, "Segmentation of speech using speaker identification," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Adelaide, South Australia), vol. 1, pp. 161–164, April 1994.

[9] R.Dunn, D. Reynolds, and T. Quatieri, "Approaches to speaker detection and tracking in conversational speech," *Digital Signal Processing*, vol. 10, nos. 1–3, pp. 93–112, January 2000.

[10] D. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Hong Kong), April 2003.

[11] B. Pellom and K. Hacioglu, "Recent improvements in the CU SONIC ASR system for noisy speech: The SPINE task", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Hong Kong), April 2003.

[12] B. Pellom, *SONIC: The University of Colorado Continuous Speech Recognizer*, University of Colorado, Technical Report TR-CSLR-2001-01, (Boulder, Colorado), March 2001.

[13] D. Reynolds, T. Quatieri, and R.Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, nos. 1–3, pp. 19–41, January 2000.

[14] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, October 1994.

[15] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, nos. 1–3, pp. 42–54, January 2000.

[3] For the various four-wire testing conditions, NIST arranges the test control files so that only same-gender tests are performed under the presumption that cross-gender tests should be easy; however, sites are not allowed to make use of this fact [3].